UNITED STATES PATENT APPLICATION

FOR

METHODS AND APPARATUSES FOR SEARCHING BOTH EXTERNAL PUBLIC DOCUMENTS AND INTERNAL PRIVATE DOCUMENTS IN RESPONSE TO A SINGLE SEARCH REQUEST

INVENTORS:

KURT W. PIERSOL
JAMEY GRAHAM

PREPARED BY:

BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN, LLP
12400 WILSHIRE BOULEVARD
SEVENTH FLOOR
LOS ANGELES, CA 90025-1026

(503) 684-6200

EXPRESS MAIL CERTIFICATE OF MAILING

"Express Mail" mailing label number: EL348719810US
Date of Deposit: September 30, 1999
I hereby certify that I am causing this paper or fee to be deposited with the United States Postal Service "Express Mail Post Office to Addressee" service on the date indicated above and that this paper or fee has been addressed to the Commissioner of Patents and Trademarks, Washington, D. C. 20231

Sharon M. Osofsky
(Typed or printed name of person mailing paper or fee)

Sharon M. Osofsky
(Signature of person mailing paper or fee)

9-30-99
(Date signed)

METHODS AND APPARATUSES FOR SEARCHING BOTH EXTERNAL PUBLIC DOCUMENTS AND INTERNAL PRIVATE DOCUMENTS IN RESPONSE TO A SINGLE SEARCH REQUEST

FIELD OF THE INVENTION

The invention relates to data processing. More specifically, the invention relates to searching and/or retrieval of public external documents and private internal documents that have been unconsciously captured in response to a single search request.

BACKGROUND OF THE INVENTION

Internet portals (or gateways) are specialized World Wide Web sites that provide a starting site for users accessing the Web. Typically, these portals provide globally useful content and searching capabilities. Portals primarily provide value by helping users find and use Web content.

Typical services offered by portal sites include a directory of Web sites, a facility to search for other sites, news, weather information, e-mail, stock quotes, phone and map information, etc. However, portals only provide the ability to search documents that have been made public by the publishers of the documents. Many documents are not available to these portals.

Many organizations have a document management policy for managing and maintaining internal private documents. Because these documents are intended to be private, they are not available to the public for searching via portals. Thus, if a searching party wishes to search both public and private documents, at least two search requests are required and at least two search results must be evaluated. Generating multiple search requests and evaluation of multiple search results can be time consuming and inefficient.

Therefore, what is needed is an improved scheme for searching public and private documents.

SUMMARY OF THE INVENTION

Methods and apparatuses for searching both external public documents and internal private documents in response to a single search request is described. A first search request is generated automatically with an electronic device in response to an original search request. The first search request to cause a search to be performed on electronic documents unconsciously captured by a local network device. The search of the electronic documents unconsciously captured is performed according to search parameters of the original search request. A second search request is generated automatically with the electronic device in response to the original search request. The second search request causes a search to be performed on electronic documents available via a network portal of an external network according to the search parameters of the original search request.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is illustrated by way of example, and not by way of limitation in the figures of the accompanying drawings in which like reference numerals refer to similar elements.

Figure 1 illustrates one embodiment of a file management system.

Figure 2A illustrates one embodiment of unconscious capture using the MIME format.

Figure 2B illustrates one embodiment of unconscious capture in an FMA environment.

Figure 2C illustrates one embodiment of the document storage process in a FMA environment.

Figure 3 illustrates one embodiment of a block diagram of a portal appliance.

Figure 4 is a flow diagram of one embodiment of a process for searching public documents and private documents in response to a single request.

Figure 5 is one embodiment of a flow diagram of a processor for capturing public content at predetermined times for unconscious capture.

DETAILED DESCRIPTION

Methods and apparatuses for searching both external public documents and internal private documents in response to a single search request is described. In the following description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the invention. It will be apparent, however, to one skilled in the art that the invention can be practiced without these specific details. In other instances, structures and devices are shown in block diagram form in order to avoid obscuring the invention.

Reference in the specification to "one embodiment" or "an embodiment" means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the invention. The appearances of the phrase "in one embodiment" in various places in the specification are not necessarily all referring to the same embodiment.

Some portions of the detailed descriptions which follow are presented in terms of algorithms and symbolic representations of operations on data bits within a computer memory. These algorithmic descriptions and representations are the means used by those skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated. It has proven convenient

at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers, or the like.

It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the following discussion, it is appreciated that throughout the description, discussions utilizing terms such as "processing" or "computing" or "calculating" or "determining" or "displaying" or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system's registers and memories into other data similarly represented as physical quantities within the computer system memories or registers or other such information storage, transmission or display devices.

The present invention also relates to apparatus for performing the operations herein. This apparatus may be specially constructed for the required purposes, or it may comprise a general purpose computer selectively activated or reconfigured by a computer program stored in the computer. Such a computer program may be stored in a computer readable storage medium, such as, but is not limited to, any type of disk including floppy disks, optical disks, CD-ROMs, and magnetic-optical disks, read-only memories (ROMs), random access memories (RAMs), EPROMs, EEPROMs, magnetic or optical cards, or any type of media suitable for storing electronic instructions, and each coupled to a computer system bus.

The algorithms and displays presented herein are not inherently related to any particular computer or other apparatus. Various general purpose systems may be used

with programs in accordance with the teachings herein, or it may prove convenient to construct more specialized apparatus to perform the required method steps. The required structure for a variety of these systems will appear from the description below. In addition, the present invention is not described with reference to any particular programming language. It will be appreciated that a variety of programming languages may be used to implement the teachings of the invention as described herein.

Methods and apparatuses for both public documents and private documents in response to a single search request are disclosed. Public documents are electronic documents that are made available to large groups of people in by the publisher of the document. An example of a public document is a World Wide Web page. Private documents are documents that have restricted access. An example of a private document is a document that generated by members of an organization and is available only to members of the organization. As described in greater detail below, a portal appliance or other device can be used to search both public and authorized private electronic documents in response to a single search request thereby improving the results of the search and/or reducing the number of searches required to find the desired material.

System Overview

Figure 1 illustrates one embodiment of a file management system. Client 110 represents a general purpose digital computer coupled to network 100. Network 100 may represent a local area network (LAN), an intranet, the Internet, or any other interconnected data path across which multiple devices may communicate. Also connected to network 100 is facsimile machine 120, copier 125, printer 130, scanner 135,

data storage device 140, server 145, file management appliance ("FMA") 150, and portal appliance ("PA") 160.

Facsimile machine 120 is connected to network 100 and represents a device capable of transmitting and receiving data such as text and images over a telephone or other communications line ("faxing"). In one embodiment, facsimile machine 120 may transmit text and images originating in printed form, or in another embodiment, facsimile machine 120 may transmit electronic data originating from any number of devices connected to network 100. Similarly, in one embodiment, facsimile machine 120 may print a hard copy of the received data, or in another embodiment, facsimile machine 120 may forward the received data to any number of devices connected to network 100.

Copier 125 represents a device capable of reproducing text and images. In one embodiment, copier 125 is a photocopier that reproduces printed text and images, whereas in another embodiment copier 125 is a photocopier that reproduces data received from any number of devices connected to network 100.

Printer 130 represents a device capable of converting electronic data into printed text and images, whereas scanner 135 represents a device capable of converting printed text and images into electronic data. In one embodiment, facsimile machine 120, photocopier 125, printer 130, and scanner 135 are each separate and distinct devices connected to network 100. In another embodiment, a multifunction device may replace any combination of these devices. Any number of devices may be omitted from or added to network 100 without parting from the spirit and scope of the present invention.

In one embodiment, data storage device 140 is also coupled to network 100. In one embodiment, data storage device 140 represents a removable storage medium such as

a CD-ROM, DVD-ROM, DVD-RAM, DVD-RW, magnetic tape or other storage medium. In an alternative embodiment, data storage device 140 represents a non-removable storage medium such as a hard or fixed disk drive. In one embodiment, data storage device 140 is an archiving device.

Server 145 represents a general purpose digital computer connected to network 100 and is configured to provide network services to other devices connected to network 100. In one embodiment, server 145 provides file sharing and printer services to network 100. In another embodiment, server 145 is a Web server that provides requested hypertext markup language (HTML) pages or files over network 100 to requesting devices. In yet another embodiment, server 145 is a server capable of providing configuration services to network 100.

FMA 150 is a file management appliance that is connected to network 100. In one embodiment, FMA 150 provides document capture and indexing services. In one embodiment, FMA 150 is a device capable of providing configuration services in addition to document capture and indexing services to network 100. In one embodiment, FMA 150 is not directly connected to any device, but rather is communicatively coupled to other devices through network 100. FMA 150 is capable of publishing its presence to other devices on network 100 using the HTTP or other protocols.

Automatic document capture (or "unconscious capture"), which is discussed more fully below, is the process by which one device, requests an archiving device, such as data storage device 140, to archive a document. In one embodiment, FMA 150 is the requesting device; however, other devices can also request archival of documents. Greater detail with respect to capture of documents that are copied, faxed, printed and other documents as well

as document management are disclosed in U.S. Patent No. 08/754,721, entitled "AUTOMATIC AND TRANSPARENT DOCUMENT ARCHIVING" filed November 21, 1996, and U.S. Patent No. 5,893,908, entitled "DOCUMENT MANAGEMENT SYSTEM," issued April 13, 1999, both of which are incorporated by reference and assigned to the corporate assignee of the present U.S. patent application.

A document may be composed of many distinct files of varying types, each representing at least the partial content of the document. A print job created on client 110 and intended for printer 130 could be captured, for example, as a thumbnail image, a postscript file, a portable document format (PDF) file, and an ASCII file containing extracted text. Additionally, FMA 150 is able to process multiple image file formats including the joint photographic experts group format (JPEG), graphics interchange format (GIF), and tagged image file format (TIFF) to name just a few. In one embodiment, each unique file type is represented by a corresponding unique file extension appended to the file's name. For example, a portable document format file may be represented as: *filename*.pdf, whereas a thumbnail image may be represented as: *filename*.thumb.

In one embodiment, FMA 150 is able to interpret compound filename extensions. For example, a thumbnail image file that contains images in a tagged image file format may be represented as *filename*.thumb.tiff. In one embodiment, FMA 150 uses the page number of the document as the filename. In such a manner, a document may be represented by multiple files located in the same directory, each representing a different page of the document as reflected by the filename. For example, "01.thumb.jpg" would represent a thumbnail image of page one in joint photographic experts group format. Similarly,

"12.thumb.tiff" would represent a thumbnail image of page twelve in tagged image file format.

FMA 150 may index data captured from various devices connected to network 100 including printer 130, facsimile machine 120, client 110 and scanner 135. In one embodiment, facsimile machine 120 captures data over a telephone line and subsequently sends at least part of the received data to FMA 150 over network 100. In another embodiment, data sent from client 110 to facsimile machine 120 over network 100 is transparently (unbeknownst to the device) captured and at least part of the data is routed to FMA 150 for indexing.

In an alternative embodiment, facsimile machine 120 is located internal to client 110 thereby eliminating the need for client 110 to send data over network 100. In such an embodiment, FMA 150 nonetheless receives at least part of the captured data. In one embodiment, FMA 150 receives bibliographic-type data extracted from the document. In one embodiment, data received from facsimile machine 120 is composed in TIFF format, whereas data received from client 110 may retain its original format upon transfer.

The FMA capture process similarly applies to other devices connected to network 100 such as scanner 135 and copier 125. In one embodiment, if optical character recognition ("OCR") is performed on a scanned or copied document, FMA 150 creates two special OCR-related files. In one embodiment, "contents.txt" and "contents.pdf" are created and used by FMA 150 to index the full text of the document and return page images as a document file respectively.

In one embodiment, FMA 150 is capable of providing the same functionality as any one or more of the devices on network 100, thereby eliminating the need for these

additional specialized devices. In one embodiment, however, FMA 150 is implemented as a thin server containing enough hardware and software to support document capture and indexing over network 100.

PA 160 is also coupled to network 100. In one embodiment, PA 160 supports searches of captured internally available, or private, documents stored, for example, on data storage device 140 as well as externally available, or public, documents available from network 170. In an alternative embodiment, the functionality of PA 160 is incorporated into FMA 150 or another device (e.g., client 110, server 145) coupled to network 100. In one embodiment network 170 is the Internet; however, network 170 can be any network of electronic devices.

In one embodiment, PA 160 operates with one or more Internet portals (e.g., yahoo.com, excite.com, go.com) to provide searching capability of external documents. Any Internet portal or any portal to an external network (e.g., a portal to a second network controlled by the organization that controls network 100) can also be used by PA 160 to provide searches of external documents. In one embodiment, the portal controls the content presented to the searching party by providing "gaps" in the search report that can be "filled" by PA 160 to present a unified search result to the searching party. In an alternative embodiment, PA 160 controls the content presented by the searching party by generating search requests to one or more portals as well as a search of data storage device 140 to search private documents. PA 160 compiles the search results and presents the results to the searching party.

Unconscious Capture

Unconscious capture is an operation in which a device (e.g., FMA 150) requests an archiving device (e.g., data storage device 140) to archive a document. In general, unconscious capture refers to FMA 150, or other device, automatically capturing documents processed by network 100 or devices coupled to network 100 without user intervention. In one embodiment, a user can optionally prevent capture of one or more documents or modify which documents are automatically captured. This may be performed by operating a selection unit or device (e.g., pressing a button).

Unconscious capture can be performed by any network entity or device. In one embodiment, unconscious capture utilizes standard Internet protocols and allows the capture of multiple files associated with a single document. In another embodiment, simultaneous capture of multiple documents is supported.

In one embodiment, a document is represented by a directory containing one metadata file and at least one data file. The actual name of the document directory is not important during unconscious capture as the name of the document is not stored as part of the directory system, but is instead stored within the metadata file. In one embodiment, the name of the document is stored in the metadata file using a document serial number. In one embodiment, the capture date is used for the name of the document directory.

In one embodiment, the capture protocol is an implementation of the Internet File Transfer Protocol (FTP). In one embodiment, documents are captured either as multipurpose Internet mail extension (MIME) files in the default FTP directory, or as subdirectories of the default directory. Other capture formats can also be used.

Figure 2A illustrates one embodiment of unconscious capture using the MIME format. A capturing device creates a MIME multi-part file, including all content files and a metadata file, 210. The capturing device then attempts to establish an anonymous FTP session with the destination device, 215. Once an FTP session is established, the capturing device determines a filename that is a unique on the destination device, 220 and attempts to transfer the file to the destination device, 225. If the transfer fails, the capturing device obtains a new filename and attempts the file transfer again. The capture is complete upon a successful file transfer, 230.

Figure 2B illustrates one embodiment of unconscious capture in an FMA environment. The capturing device establishes an anonymous FTP session with the destination device, 235. Once the FTP session is established, the capturing device determines what it assumes to be a unique directory name on the destination device, 240. Once a directory name is determined, the capturing device attempts to create a directory with that name on the destination device, 245. If the attempt to create the directory is unsuccessful, whether due to a duplicate directory name or otherwise, the capturing device determines another directory name and attempts to create the directory again.

If, however, the capturing device successfully creates the directory on the destination device, 250, the capturing device then copies the content file or files to the newly created directory, 255. The capturing device also creates a metadata file, 260, which is then sent to the FMA device, 265 to complete the process.

Figure 2C illustrates one embodiment of the document storage process in a FMA environment. In one embodiment, the document directory is represented by "yyyy/mm/dd" where *yyyy* represents the year in which the document was created, *mm*

represents the ordinal month in which the document was created, and *dd* represents the day of the month in which the document was created. Other date formats and/or storage ordering can also be used.

~~During the document storage process, the FMA creates appropriate directories,~~ moves the document to the appropriate directory, and updates the master list. The metadata file of the document to be stored is accessed and information from its "Capture date" field is retrieved, 270. If the document's "Capture date" or even the metadata file does not exist, then the current system time is obtained and used as the document's "Capture date," 274. If, however, the document's "Capture date" does exist, the system ~~determines whether an appropriately named directory exists~~.

The system determines whether a directory exists as reflected by the appropriate year, 276. If a directory reflecting the appropriate year does not exist, the system creates such a directory, 278. If a directory reflecting the appropriate year does exist, the system then checks whether a directory reflecting the appropriate month exists within that year directory, 280. If the appropriate month directory does not exist within the year directory, the system creates a month directory within the year directory, 282. If the appropriate year and month directories exist, the system finally checks whether the appropriate day directory exists within the nested year/month directory, 284.

If the day directory does not exist, the system creates the appropriate day directory within the year/month directory, 286. If a directory reflecting the appropriate year, month and day already exists, the system creates a new document directory name into which the document will be stored. In one embodiment, the system generates a four-digit random number that gets appended to the end of the existing document directory

name, 288. Once a unique document directory name is established, 286 and 288, the document is moved to that directory, 290 and the master document list is updated to reflect the document's new location, 292.

Overview of a Portal Appliance

Figure 3 illustrates one embodiment of a block diagram of a portal appliance. PA 160 includes bus 310 or other communication device to communicate information, and processor 320 coupled to bus 310 to process information. While PA 160 is illustrated with a single processor, PA 160 can include multiple processors and/or co-processors. PA 160 further includes random access memory (RAM) or other dynamic storage device 350 (referred to as main memory), coupled to bus 310 to store information and instructions to be executed by processor 320. Main memory 350 also can be used to store temporary variables or other intermediate information during execution of instructions by processor 320.

PA 160 also includes read only memory (ROM) and/or other static storage device 330 coupled to bus 310 to store static information and instructions for processor 320. Storage device 370 is coupled to bus 310 to store information and instructions. Storage device 370 such as a magnetic disk or optical disc and corresponding drive can be coupled to PA 160.

PA 160 can also be coupled via bus 310 to I/O devices 360, such as a cathode ray tube (CRT) or liquid crystal display (LCD), to display information to a user, and alphanumeric input device to communicate information and command selections to processor 320. Another type of I/O device is a cursor control, such as a mouse, a trackball, or cursor direction keys to communicate direction information and command selections to

processor 320 and to control cursor movement on the display. Additional and/or different I/O devices can also be coupled to bus 310.

Network interface 340 provides an interface between PA 160 and network 170. Similarly, network interface 345 provides an interface between PA 160 and network 100. In one embodiment, network interface 340 and network interface 345 are network interface cards (NICs), which are known in the art; however, any interface that can provide PA 160 with access to multiple networks can be used.

One embodiment of the invention is related to the use of PA 160 to perform searches on both network 100 and network 170. According to one embodiment, the searches are performed by PA 160 in response to processor 320 executing sequences of instructions contained in main memory 350. Instructions are provided to main memory 350 from a storage device, such as magnetic disk, a read-only memory (ROM) integrated circuit (IC), CD-ROM, DVD, via a remote connection (e.g., over a network), etc. In alternative embodiments, hard-wired circuitry can be used in place of or in combination with software instructions to implement the present invention. Thus, the present invention is not limited to any specific combination of hardware circuitry and software instructions.

Document Searching

Figure 4 is a flow diagram of one embodiment of a process for searching public documents and private documents in response to a single request. The portal appliance receives a search request, 410. The search request can be in the form of a boolean text search, a plain language text search, or any other appropriate format.

The portal appliance sends the search request to one or more portals, 420. In one embodiment, the portal appliance performs any necessary request reformating such that the portal search requests are recognized by the portal receiving the search request. Similarly, the portal appliance sends the search request to a networked FMA or other device to perform searches on unconsciously captured documents, 430. The portal search(es) and the captured document search can be performed in parallel or sequentially.

Results from the portal search are received, 440. Similarly, results from the captured document search are received, 450. The search results can be received in parallel or sequentially. The portal appliance combines the search results 460.

In one embodiment, the portal search report indicates where the captured document search report is to be inserted. This embodiment provides the portal with control of the style and content of the search report. If the search results are provided in Hypertext Markup Language (HTML) format, for example, an anchor (<A>) tag can be used to indicate where the captured document search report is to be inserted. For example

<A HREF="search:foo" MODE="Table" WIDTH = 300>

indicates that the tag should be replaced with a captured document search for the word "foo" presented in a table with a width of 300 pixels. Of course, any other format can also be used.

In an alternative embodiment, the portal appliance controls the style and content of the search report by issuing search reports to one or more portals and to the FMA to search the captured documents. In one embodiment, the portal appliance generates Hypertext Transfer Protocol (HTTP) requests to the portals, which can be, for example, Common Gateway Interface (CGI) programs that perform the requested searches. The portal

appliance receives the multiple search requests and combines the search requests into an appropriate format.

The portal appliance outputs the search results in response the combination of the search results, 470. In one embodiment, the search results are presented as an HTML document having links to the documents identified by the searches. Other formats, for example, Extensible Markup Language (XML) can also be used. If a user selects one of the links the document is retrieved from either a private source coupled to a network belonging to the organization or from an external public source.

Thus, a portal appliance or other device can integrate content from a portal with content from unconsciously captured documents to provide a searching party with a unified search result. The unified search result provided by the portal appliance allows a searching party to search both public documents published by other parties and internal documents accessible by the searching party in response to a single search without requiring a the publisher of the private documents to publicly release the content of the private documents.

In one embodiment, the search report output by the portal appliance includes advertising that is based on the search performed. The search terms can be used to determine the advertising to be displayed to the searching party. Advertising can also be provided by the portal appliance based on the search terms or other information, for example, an internal user profile. By providing information for selection of advertising and displaying the advertising the organization controlling the portal appliance can receive advertising revenue.

Figure 5 is one embodiment of a flow diagram of a process for capturing public content at predetermined times for unconscious capture. In one embodiment, the portal

appliance can retrieve information at predetermined times. The retrieved information can be used, for example, to populate a database with information provided but not archived by a portal (e.g., stock quotes, press releases, news). The portal appliance retrieves the information and the FMA causes the information to be captured and archived. The archived information can the be accessed and/or searched at a later time.

~~The portal appliance waits for a predetermined time for retrieving information, 510.~~ Predetermined content is retrieved in response to a request at the predetermined time, 520. The content can be retrieved by the portal appliance "pulling" the content, for example, in the form of one or more HTTP requests initiated by the portal appliance. The content can also be "pushed" by an external portal to the portal appliance, for example, with an HTTP or FTP operation. The content is captured by the portal appliance, 530. The content is archived by, for example, the FMA, 440. If additional retrievals are scheduled, 550, the ~~process is repeated~~.

In the foregoing specification, the invention has been described with reference to specific embodiments thereof. It will, however, be evident that various modifications and changes can be made thereto without departing from the broader spirit and scope of the invention. The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense.